Docket: PH01-00-10

METHOD AND APPARATUS FOR ALLEVIATING TRAFFIC CONGESTION IN A COMPUTER NETWORK

FIELD OF THE INVENTION

[0001] The present invention relates generally to a method and apparatus for alleviating congestion in a computer network and more generally to a method and apparatus for alleviating congestion in a computer network that employs multiple layers of protocols.

Background of the Invention

[0002] To facilitate an understanding of the present invention, an overview of network protocols and devices will be presented below.

COMPUTER NETWORK PROTOCOLS

[0003] At the heart of any computer network is a communication protocol. A protocol is a set of conventions or rules that govern the transfer of data between computer devices. The simplest protocols define only a hardware configuration, while more complex protocols define timing, data formats, error detection and correction techniques and software structures.

[0004] Computer networks almost universally employ multiple layers of protocols. A low-level physical layer protocol assures the transmission and reception of a data stream between two devices. Data packets are constructed in a data link layer. Over the physical layer, a network and transport layer protocol governs transmission of data through the network, thereby ensuring end-to end reliable data delivery.

[0005] As computer networks have developed, various approaches have been used in the choice of communication medium, network topology, message format, protocols for channel access, and so forth. Some of these approaches have emerged as de facto standards, but there is still no single standard for network communication. However, a model for network architectures has been proposed and widely accepted. It is known as the International Standards Organization (ISO)

Open Systems Interconnection (OSI) reference model. The OSI reference model is not itself a network architecture. Rather it specifies a hierarchy of protocol layers and defines the function of each layer in the network. Each layer in one computer of the network carries on a conversation with the corresponding layer in another computer with which communication is taking place, in accordance with a protocol defining the rules of this communication. In reality, information is transferred down from layer to layer in one computer, then through the channel medium and back up the successive layers of the other computer. However, for purposes of design of the various layers and understanding their functions, it is easier to consider each of the layers as communicating with its counterpart at the same level, in a "horizontal" direction.

[0006] The lowest layer defined by the OSI model is called the physical layer, and is concerned with transmitting raw data bits over the communication channel. Design of the physical layer involves issues of electrical, mechanical or optical engineering, depending on the medium used for the communication channel. The layer next to the physical layer is called the data link layer. The main task of the data link layer is to transform the physical layer, which interfaces directly with the channel medium, into a communication link that appears error-free to the next layer above, known as the network layer. The data link layer performs such functions as structuring data into packets or frames, and attaching control information to the packets or frames, such as checksums for error detection, and packet numbers.

[0007] Although the data link layer is primarily independent of the nature of the physical transmission medium, certain aspects of the data link layer function are more dependent on the transmission medium. For this reason, the data link layer in some network architectures is divided into two sublayers: a logical link control sublayer, which performs all medium-independent functions of the data link layer, and a media access control (MAC) sublayer. This sublayer determines which station should get access to the communication channel when there are conflicting requests for access. The functions of the MAC layer are more likely to be dependent on the nature of the transmission medium.

[0008] The Internet is a protocol-specific collection of networks, including Arpanet, NSFnet, regional networks such as NYsernet, local networks at a number of university and research institutions, and a number of military networks. The ID protocols used by the Internet are generally referred to as the TCP/IP suite of protocols. TCP/IP deviates from the seven-layer OSI model in that it has five layers. These five layers are combinations and derivatives of the seven-layer OSI model. A rough correspondence between the seven layers of the OSI model and TCP/IP is shown in FIG. 1.

[0009] The protocols provide a set of services that permit users to communicate with each other across the entire Internet. The specific services that these protocols provide are not important to the present invention, but include file transfer, remote login, remote execution, remote printing, computer mail, and access to network file systems.

[0010] The basic function of the Transmission Control Protocol (TCP) is to make sure that commands and messages from an application protocol, such as computer mail, are sent to their desired destinations. TCP keeps track of what is sent, and retransmits anything that does not get to its destination correctly. If any message is too long to be sent as one "datagram," TCP will split it into multiple datagrams and makes sure that they all arrive correctly and are reassembled for the application program at the receiving end. Since these functions are needed for many applications, they are collected into a separate protocol (TCP) rather than being part of each application. TCP is implemented in the transport layer of the OSI reference model.

[0011] The Internet Protocol (IP) is implemented in the network layer of the OSI reference model, and provides a basic service to TCP: delivering datagrams to their destinations. TCP simply hands IP a datagram with an intended destination; IP is unaware of any relationship between successive datagrams, and merely handles routing of each datagram to its destination. If the destination is a station connected to a different LAN, the IP makes use of routers to forward the message.

[0012] IP routing specifies that IP datagrams travel through internetworks one hop at a time (next hop routing) based on the destination address in the IP header. The entire route is not known at the outset of the journey. Instead, at each stop, the next destination (or next hop) is calculated by matching the destination address within the datagram's IP header with an entry in the current node's (typically but not always a router) routing table.

[0013] Each node's involvement in the routing process consists only of forwarding packets based on internal information resident in the router, regardless of whether the packets get to their final destination. To extend this explanation a step further, IP routing does not alter the original datagram. In particular, the datagram source and destination addresses remain unaltered. The IP header always specifies the IP address of the original source and the IP address of the ultimate destination.

NETWORK LEVEL DEVICES

[0014] The end systems (e.g., computers or printers) which form a computer network are interconnected by devices known as routers. Each end system is attached to one of the network's routers and each router is responsible for forwarding communications to and from its attached end systems. A router operates at the network layer level and implements the IP protocol. When a router receives a data packet, it reads the data packet's destination address from the data packet header, and then transmits the data packet on the link leading most directly to the data packet's destination. Along the path from source to destination, a data packet may be transmitted along several links and pass through several routers, each router on the path reading the data packet header and forwarding the data packet accordingly.

[0015] To determine how user data packets should be forwarded, each router typically knows the locations of the network's end systems (i.e., which routers are responsible for which end systems), the nature of the connections between the routers, and the states (e.g., operative or inoperative) of the links forming those connections. Using this information, each router can compute effective routes

through the network and avoid, for example, faulty links or routers. A procedure for performing these tasks is generally known as a "routing algorithm."

[0016] A router also performs protocol translation. One example is at layers 1 and 2 of the TCP/IP model. If the datagram arrives via an Ethernet interface and is destined to exit on a serial line, for example, the router will strip off the Ethernet header and trailer, and substitute the appropriate header and trailer for the specific network media, such as SMDS, by way of example.

PHYSICAL LAYER DEVICES

[0017] The TCP/IP protocol suite may operate over many different physical layer technologies. For example, rapidly growing use is being made of optical fiber as the physical transmission medium through which optical signals travel. Optical communication systems may be configured to carry an optical channel of a single wavelength over one or more optical waveguides. To convey information from plural sources time-division multiplexing (TDM) is sometimes employed. In time-division multiplexing, a particular time slot is assigned to each information source, the complete signal being constructed from the signal portion collected from each time slot. While this is a useful technique for carrying plural information sources on a single channel, its capacity is limited by fiber dispersion and the need to generate high peak power pulses.

[0018] An alternative multiplexing technique that may be employed to increase capacity is wavelength division multiplexing (WDM). A WDM system employs plural optical signal channels, each channel being assigned a particular channel wavelength. In a WDM system, optical signal channels are generated, multiplexed to form an optical signal comprised of the individual optical signal channels, transmitted over a single waveguide, and demultiplexed such that each channel wavelength is individually routed to a designated receiver.

[0019] The physical layer device corresponding to a router (which, as mentioned, is a network or IP level device) is an optical switch. In a WDM system, an optical switch allows different wavelength channels to be directed along different paths in the network. Optical switches may be fixed wavelength-

dependent elements in which a given wavelength is always routed along a given path. More flexible optical switches are reconfigurable elements that can dynamically change the path along which a given wavelength is routed. Examples of a fixed optical switch include Add/Drop Multiplexers (OADM) and Optical Cross-Connects (OXC) such as disclosed in U.S. Patent Nos. 5,504,827, 5,612,805, and 5,959,749, while general OXC switching architecture is reviewed by E. Murphy in chapter 10 of Optical Fiber Telecommunications IIIB, edited by T. Koch and I. Kaminow, for example. An example of a more flexible reconfigurable optical switch is disclosed in U.S. Application Serial No. [PH01-00-01].

NETWORK CONGESTION

[0020] Data networks are typically designed and installed with a particular traffic pattern in mind. This traffic pattern dictates the size of routers, switches, repeaters, and other equipment, which are used to accommodate the maximum expected capacity at each node in the network. However, traffic flowing in a data network changes dramatically over time and, as a result, the physical location and size of routers and switching engines often do not match capacity requirements. Currently, networks and their management systems require that capacity upgrades be performed locally where a bottleneck occurs. This is typically done by physically adding switching capacity to the node in question, including the addition of repeaters and the like. In addition to being an expensive, time-consuming, and a slow-to-respond solution, such an approach fails to take advantage of other portions of the network, which at any given time may have excess capacity.

[0021] Accordingly, it would be desirable to alleviate network congestion that arises in one portion of a network by taking advantage of available resources in another portion of the network.

Summary of the Invention

[0022] In accordance with the present invention, a method is provided for alleviating congestion in a computer network. The network includes a plurality of nodes interconnected by communication links, which communicate in accordance with at least an optical layer protocol and a second protocol layer. The method begins by receiving a signal in accordance with the optical layer protocol and determining if local forwarding capacity is available to forward the signal in accordance with the second protocol layer. If the local forwarding capacity is unavailable, the method continues by forwarding the signal in accordance with the optical layer protocol to another node having excess capacity so that the other node can forward the signal in accordance with the second protocol layer.

[0023] In accordance with one aspect of the invention, the second protocol layer includes a network layer protocol.

[0024] In accordance with another aspect of the invention, the optical layer protocol includes wavelength division multiplexing.

[0025] In accordance with yet another aspect of the invention, the network layer protocol is an Internet protocol.

[0026] In accordance with another aspect of the invention, the signal is received in accordance with the first protocol layer by an optical switch.

[0027] In accordance with another aspect of the invention, the optical switch includes an add/drop multiplexer or an optical cross-connect.

[0028] In accordance with another aspect of the invention, the optical switch is a reconfigurable optical element.

[0029] In accordance with yet another aspect of the invention, the local forwarding capacity is provided by a packet switching element such as an IP router.

[0030] In accordance with another aspect of the invention, a communication system is provided which includes a plurality of network nodes for receiving and forwarding data traffic. Each of said nodes has a prescribed optical switching capacity and a prescribed packet switching capacity that is less than the prescribed optical switching capacity. At least one communication link interconnects the network nodes. The system also includes a structure, embodied in hardware,

software, or a combination of both, and which is associated with a given network node, for distributing traffic to be packet switched to another one of the network nodes if the prescribed packet switching capacity of the given network node is exceeded.

[0031] In accordance with another aspect of the invention, each of the plurality of network nodes includes traffic distributing structure.

[0032] In accordance with yet another aspect of the invention, the network nodes include an optical switch that provides the prescribed optical switching capacity.

[0033] In accordance with another aspect of the invention, the optical switch includes the traffic distributing structure.

Brief Description of the Drawings

[0034] FIG. 1 is a comparative diagram of the International Standards Organization (ISO) Open System Interconnection (OSI) model for network architectures and a commonly used TCP/IP protocol.

[0035] FIG. 2 shows a simplified block diagram of an exemplary computer network that includes multiple nodes and end systems interconnected by communication links.

[0036] FIG. 3 shows a simplified schematic diagram of a network node constructed in accordance with the present invention.

[0037] FIG. 4 shows a simplified block diagram of an exemplary computer network that includes multiple ones of the inventive nodes shown in FIG. 3.

Detailed Description

[0038] The present invention addresses the problem that arises when there are local pockets of congestion in a network, which create bottlenecks of traffic that need to be routed, while at the same time other equipment is available to transfer the congested traffic to underutilized resources in the network that can more readily route the traffic. For example, the invention is applicable to optical data networks in which optical switches and high-capacity optical transmission lines allow large

amounts of data to be forwarded to unused resources elsewhere in the network such as underutilized packet switching elements, e.g., IP routers.

[0039] FIG. 2 shows an exemplary computer network that includes end units interconnected by nodes. For purposes of illustration only the computer network will be depicted as a WDM optical communication system over which the TCP/IP suite of protocols is implemented. However, it should be noted that the present invention is equally applicable to other network protocols and technologies and is not limited to an all-optical transmission medium or IP packet switching.

[0040] As shown, the end units are the origination and destination points of traffic and will be referred to herein as customer premises equipment (CPE). Nodes 101, 102, and 103 are connected to nodes 1051-1053, 1071-1073, and 1091-1093, respectively. The transmission paths connecting the nodes and CPEs are assumed to be all-optical paths in which WDM signals travel. Transmission paths connecting one node to another node will be referred to as transmission interfaces and transmission paths connecting one CPE to another CPE will be referred to as connection interfaces. For example, path 115 is a transmission interface and path 116 is a connection interface.

[0041] Each node includes the equipment necessary to perform optical switching of the channels and packet switching of individual data packets. Accordingly, each node includes a router and an optical switch. A simplified schematic diagram of a node 300 is shown in FIG. 3. In this illustrative embodiment of the invention the optical switch 310 is depicted as an optical cross-connect and the router 312 is depicted as an IP router. As shown, the node 300 includes a plurality of input/output ports 33201-3204 and 3301-3304 for receiving the WDM optical signals from the optical transmission paths. Once optical cross-connect 310 receives a WDM optical signal, it may forward the signal to the router via interface 314, where it will undergo optical to electrical conversion. As previously described in more detail, the router 312 assembles the data packets and reads their destination address to determine the subsequent nodes to which the packets are to be forwarded. The data packets then undergo electrical to optical conversion in

interface 314 and are directed onto the appropriate channel so that they can be forwarded to the appropriate nodes by the optical switch 310.

[0042] As previously mentioned, a situation may arise in which the amount of traffic received by a given node exceeds the throughput capacity of the router. Since the throughput capacity of the optical switch 310 is generally far greater than the throughput capacity of the router 312, the throughput capacity of the router 312 will generally be exceeded before that of the optical switch 310. That is, at any given time the optical switch 310 may retain the capacity to perform optical switching while its associated router 312 cannot perform any additional packet switching. In such a case, in accordance with the present invention, instead of forwarding the WDM optical signal to the router 312, the optical cross-connect 310 can forward the WDM signal directly to one of its output ports so that it is transmitted to another node. In a preferred embodiment of the invention, the optical cross-connect 310 forwards the WDM signal to another node having a router with excess throughput capacity.

[0043] FIG. 4 shows a network of three of the inventive nodes 401-403 of the type shown in FIG. 4. In this network two signals are assumed to originate from a CPE (not shown) that is in optical communication with node 401. If wavelength division multiplexing is employed, the two signals are located on different wavelength channels. One signal (shown by a dashed line in FIG. 4) is forwarded from the optical cross-connect 405 of node 401 to the router 407 of node 401. After receiving this signal the throughput capacity of the router 407 is assumed to be exceeded. In this case the second signal (shown by a dotted line) is forwarded by the optical cross-connect 405, via one of its output ports, to the optical cross-connect 409 of node 403, whose router 411 is currently underutilized. After converting the second signal to an electrical signal, the router 411 assembles the data packets and reads their destination address to determine the subsequent nodes to which the packets are to be forwarded. The data packets then undergo an electrical to optical conversion in interface 413 and are directed onto the appropriate channel so that they can be forwarded to the appropriate nodes by the optical switch 409. Accordingly, the otherwise unused capacity of node 403 is used

when traffic that requires packet switching overburdens the router 411 of node 401. Consequently, with this method of distributing packet switching among different nodes, a given network is capable, on average, of handling a higher traffic load than it could in a conventional network in which the same node performs optical switching and routing of a given signal.

[0044] Of course, more generally, the present invention is applicable to any computer network having two or more protocol layers of switching in which one layer of switching can be distributed among different nodes.